1H4-NFC-01a-1

テキストデータマイニングのための統合環境TETDMによる利用者支援

User Support in Text Data Mining with Total Environment for Text Data Mining(TETDM)

砂山 渡[*1] Wataru Sunayama
高間 康史[*2] Yasufumi Takama
西原 陽子[*3] Yoko Nishihara
徳永 秀和[*4] Hidekazu Tokunaga
串間 宗夫[*5] Muneo Kushima
阿部 秀尚[*6] Hidenao Abe
梶並 知記[*7] Tomoki Kajinami
松下 光範[*8] Mitsunori Matsushita
ダヌシカ ボレガラ[*9] Danushka Bollegala
佐賀 亮介[*10] Ryosuke Saga
河原 吉伸[*11] Yoshinobu Kawahara
川本 佳代[*1] Kayo Kawamoto

[*1]広島市立大学大学院情報科学研究科 Graduate School of Information Sciences, Hiroshima City University
[*2]首都大学東京システムデザイン学部 Faculty of System Design, Tokyo Metropolitan University
[*3]立命館大学情報理工学部 College of Information Science and Engineering, Ritsumeikan University
[*4]香川高等専門学校 Kagawa National College of Technology
[*5]宮崎大学医学部附属病院医療情報部 Medical Informatics, University of Miyazaki Hospital
[*6]文教大学情報学部 Faculty of Information and Communications, Bunkyo University
[*7]神奈川工科大学情報学部 Faculty of Information Technology, Kanagawa Institute of Technology
[*8]関西大学総合情報学部 Faculty of Informatics, Kansai University
[*9]School of Electrical Engineering, Electronics, and Computer Science, Liverpool University
[*10]大阪府立大学工学研究科 Graduate School of Engineering, Osaka Prefecture University
[*11]大阪大学産業科学研究所 The Institute of Scientific and Industrial Research, Osaka University

In this challenge, we develop and distribute an integrated environment to flexibly combine multiple text mining techniques. Text mining techniques include numerous tasks such as salient sentence extraction, keyword extraction, topic extraction, textual coherence evaluation, multi-document summarization, and text clustering. Although tools that individually perform one or more of the above-mentioned tasks exist, it is difficult to integrate and activate multiple tools for a particular task. We attempt to provide the flexibility to integrate numerous tools that exist in the community in our proposed text mining environment. Users can use a customized version of the proposed text mining environment for their specific tasks, thereby concentrating solely on their creative work.

1. はじめに

人工知能学会全国大会の近未来チャレンジテーマとして進められている TETDM（Total Environment for Text Data Mining: テトディーエム）[*1]は，複数のテキストマイニング技術を柔軟に組み合わせて使える統合環境を構築し，社会的創造的活動を支援できる環境としての提供を目指している [砂山 14a]. 近未来チャレンジでは，5 年以内の目標達成を必要条件としており，現在 4 年目 (2014 年の全国大会で 5 年目に入る) の TETDM は，統合環境の正式公開を 2015 年 4 月と目標を定めて進めている（図 1）.

これまでに，複数のモジュールを統合的に扱う環境の枠組み作りを行い，2011 年 12 月の試用α版の公開を皮切りに，2014 年 3 月現在でバージョン 0.55 までを公開している．統合環境ならびに必要な情報の公開は，TETDM サイト上で行っており，継続的にアクセスとダウンロードがなされている．

しかし，テキストマイニングツールとしての使い勝手は，これまで十分に整備されていなかった．また幅広い利用者に TETDM を利用してもらうことを目指す本チャレンジにおいては，テキストマイニングの初心者が TETDM を使いこなすスキルを身につける道筋を用意する必要があった．

そこで本稿では，2. で TETDM の構成と利用者支援に関するこの一年間で実装された内容について述べる．3. で，利用者に対して実践した内容について述べる．4. で関連研究を述べ，5. で本稿を締めくくる．

連絡先: 砂山渡，広島市立大学大学院情報科学研究科，731-3194 広島市安佐南区大塚東 3-4-1，TEL082-830-1705

*1 TETDM サイト：http://tetdm.jp/

2. TETDM 統合環境の構成

本章では，TETDM 統合環境の構成と，利用者支援のために新たに実装した機能について述べる．TETDM は，あるテキストを入力したときに，その分析処理を行う処理モジュールと，処理結果を可視化して出力する可視化モジュールを複数備えた環境となっている（図 2）．またこれらのモジュール群は，単独の開発者によって提供されるものではなく，任意の開発者

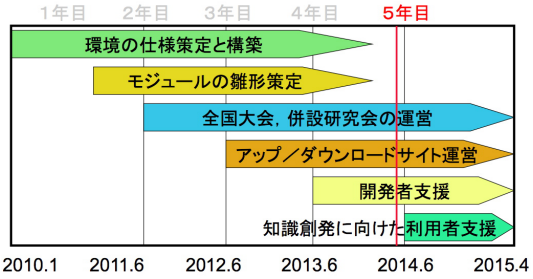

図 1: 近未来チャレンジ TETDM のスケジュール

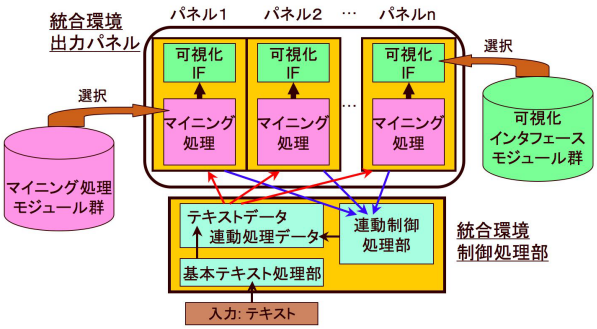

図 2: TETDM 統合環境構成図

がモジュールを作成して利用することができる．

TETDM は，独立した複数のパネル内に，処理モジュールと可視化モジュールを 1 つずつペアとしてセットすることで動作する．各パネルにセットされる異なる開発者によって作成されたモジュールが，それぞれ独立に動作するだけではなく，それらを協調的に連動させることができる点が TETDM の特徴となっている．以下で各部の説明を述べる*2．

2.1 入力：テキスト

TETDM に入力されたテキストは「セグメント」「文」「単語」の 3 つに分割して扱われる．「単語」へ区切る際は，形態素解析器を用いて単語に分割する．この際，指定した品詞の単語だけを，キーワードとして取り扱う．「文」に区切る方法は，テキスト中の句点記号（「。」や「.」）をもとに分割する．「セグメント」に区切る方法は，特定の文字列をテキスト中に挿入し，その文字列をもとに分割する．その後，単語の出現情報や頻度情報の計算，キーワードやセグメント間の関連度計算を行った結果をデータ構造に格納し，このテキストデータをもとに各モジュールが処理を行う．

2.2 マイニング処理モジュール

マイニング処理モジュールは，統合環境内のテキストデータをもとに，テキストの理解や分析に役立つ情報をテキストから抽出する．またマイニングという言葉にこだわることなく，テキストに何らかの処理を施すモジュールも対象とする．現在までに 30 以上の処理モジュールが作成，公開されている．マイニング処理モジュールの処理結果は，可視化インタフェースモジュールに渡されて出力される．

*2 構成の詳細は [砂山 14a] を参照していただきたい．

表 1: TETDM モジュールのツールタイプ

タイプ名	モジュールの定義
シンプル	データのやりとりが不要で単独で完結する
プリミティブ	一種類のデータ型のデータを 1 つだけ渡す，または受け取る
セミプリミティブ	複数のデータを渡す／受け取るが，指定モジュール以外との組合せでも動作する
特殊	データに特定の意味があり，指定モジュール以外との組合せでは動作しない

2.3 出力：可視化インタフェースモジュール

可視化インタフェースモジュールは，マイニング処理モジュールによる出力結果を可視化する*3．可視化インタフェースモジュールでは，入力として受け取れるデータ型（boolean, int, double, String 型とその一次元配列，二次元配列 (String 型以外) の 11 種類）を定め*4，そのデータの意味を表す整数型変数との組合せにより，マイニング処理モジュールからデータを受け取ることができる．現在までに 30 以上の可視化モジュールが作成，公開されている．

2.4 TETDM の利用者支援に関する実装

利用者支援のために，以下の機能を新たに実装した*5．

1. 初心者のためのチュートリアル
2. 部分テキストに対する分析
3. ツールタイプの設定
4. データ型コンバート
5. 比較分析用機能の実装

1. は，初心者が TETDM の用語や基本的な使い方をスムーズに覚えられるよう，ゲーム的に課題をクリアしていくチュートリアルを実装した．現在の実装内容は，初心者が基本的な操作を覚え，存在するモジュールを一通り試用するまでとなっているが，今後より実践的なマイニングスキルが身につけられる課題を実装していく予定となっている．

2. は，TETDM に入力したすべてのテキストに対する結果だけではなく，一部の文あるいは段落（セグメント）のみに対する結果を出力する機能を設けた．これは実際のテキスト分析においては，分析対象となるテキスト量が増えるにつれ，全体の結果を閲覧することが困難になり，ユーザの観点に基づいて情報を絞り込んだ結果に対してマイニングを行う必要性が増すと考えたことによる．

3. は，TETDM では，処理モジュールと可視化モジュールを組み合わせて利用する必要があるが，モジュールの数が増したときに，どのモジュールが組み合わせ可能かを判定しやすくすること，ならびに組合せの可能性を高めるために設定した．ツールタイプは表 1 に示す 4 つを定めており，このうち，「シンプル」と「プリミティブ」に該当する処理モジュールと可視化モジュールは，4. で説明するデータ型コンバートの機能と合わせることで，任意に組み合わせることが可能となっている．

*3 マイニング処理モジュールの結果によらず，統合環境のテキストデータを可視化するモジュールであっても良い．

*4 TETDM は Java で実装されている．

*5 本節で述べる一部の機能は，本稿作成時のバージョン (0.55) に含まれていないものもあるが，順次含めていく予定となっている．

表 2: データ型コンバートのルール

変換前＼後	boolean	int	double	String
boolean	-	0 か 1 に	0.0 か 1.0 に	文字列に
int	0 以外 true	-	(double) に	文字列に
double	0 以外 true	(int) に	-	文字列に
String	“”以外 true	文字数	文字数	-

4. は，可視化モジュールが受け取りを指定するデータ型のデータ以外でも，受け取れるデータ型に自動的に変換してデータの受け渡しを可能にする機能を実装した．最も単純な例は，整数型の配列を実数型の配列として扱うことやその逆で，文字列の場合は文字数にして数値化する（表 2）．

5. は，1. の部分テキストが複数生成されたときに，その部分テキスト間，あるいは全テキストと部分テキストの比較を可能にする実装を行った．

3. TETDM の利用者支援に関する分析評価

本章では，この一年間で利用者支援に関して行われた分析評価の概要について述べる（詳細は，他稿を参照していただきたい）．

3.1 TETDM を用いたテキスト分析初心者の観察実験

本節では，テキスト分析の初心者が TETDM を用いて，どのように分析を行うかを観察した実験 [井須 14] について述べる．

実験参加者は，情報学を専攻する大学院生 2 名（男性 2 名），大学生 2 名（男性 1 名，女性 1 名），社会人 1 名（女性 1 名）の計 5 名とした．実験に際して，予め参加者に TETDM の機能と操作方法について説明を行った．

実験課題は,「小説『ヴィヨンの妻（太宰治著）』*6 を TETDM（Ver.0.53）を用いて分析した上で登場人物を発見し，それぞれについて説明すること」とし，時間制限なしで考えがまとまるまで続けてもらった．

実験の様子とインタビューの結果から，以下の 4 つの点が初心者の分析を妨げていることがわかった．

1. 以前に使用したツールを把握できない
2. 可視化ツールの名称から可視化結果が想像しづらい
3. ツールが行う処理の進行状況が示されない
4. ツールの種類が多く，ツールの選択に手間取る

今後これらのデザイン指針を踏まえて TETDM を改良していくことで，初心者を含むユーザが，利用に戸惑うことがない環境を構築していきたい．

3.2 TETDM の利用者用チュートリアルの実装

本節では，初心者を含む TETDM のユーザが，利用方法を学ぶために実装したチュートリアルと，その効果を確認した実験 [中垣内 14] について述べる．TETDM を数回使ったことがあるか，全く使ったことがない成人 15 名に，チュートリアルを使用してもらい，その効果を確認する実験を行った．

事前テストと事後テストはそれぞれ 21 点満点で，事前テストの平均点は 6.8 点，事後テストの平均点は 12.9 点となり，後者の方が有意に高くなった (一対の標本による差の検定：$t(14) = 5.97, p < .001$)．このことから，本チュートリアルは一通りの処理ツールと可視化ツールの使い方に関する知識を習得させ，それらを生かして各自の目的に応じたテキストマイニングを試みるスキルを身につけさせる上で，一定の効果があった．

3.3 TETDM を用いたインタラクティブクラスタリングにおける対制約付与方法の比較

本節では，一部モジュールの交換により，比較対象システム間での共通性を高く維持したまま比較実験が可能という TETDM の特徴を活かして，インタラクティブクラスタリングにおける対制約（同一クラスタに入るオブジェクト対の指定）の指定方法に関して，制約一括生成手法と逐次指定手法の比較を行った結果 [北村 14] について述べる．

20 名の実験協力者に提案インタフェースを利用し，新聞記事をジャンルごとに分類してもらう実験を行った．協力者を 10 名ずつの 2 グループに分け，それぞれ制約一括生成手法，逐次指定手法を用いて作業を行ってもらった．

平均クラスタリング回数は一括生成手法の方が少ない一方，平均作業時間は逐次指定手法の方が短くなった．また，付与された制約対数は両手法で大差ない結果が得られた．このような条件を揃えたインタフェース比較実験に，TETDM は適していると考える．

3.4 TETDM による Exploratory Search

本節では，3.2 で述べたチュートリアルを行った利用者が，既存のツールを利用して Exploratory Search を行った結果 [徳永 14] について述べる．

環境問題をテーマとしたレポート作成のために，コンセプトマップを作成しながら情報を収集してもらう実験を行った．被験者は香川高専専攻科生 2 名とした．

被験者は TETDM の環境を用いることで，インタラクティブに知識の修得と再分析を複数のパネルを連動させながら行うことができた．特に，観点となる単語の選択と知識の修得を繰り返す際に，観点語に関する Web ページの要約文と，観点語に関する文のハイライト表示を，インタラクティブに確認できたことが，効率的な知識修得に役立てられたと考えられる．被験者 B は，被験者 A に比べて要約文をより活用するなど，TETDM の活用方法に差があるとともに，また被験者 A とは異なる「大気汚染」や「中国」などの情報を収集しており，ユーザの興味に応じた柔軟な情報収集を実現できていた．

3.5 TETDM を用いた知識創発の枠組みと実装

本節では，近未来チャレンジ TETDM の最終目標の 1 つとなっている知識創発支援に向けて，想定する枠組み，行った実装と実験 [砂山 14b] について述べる．

本研究では知識創発を,「解釈（事実に対する意味づけ）を積み重ねることで，汎用性がある新たな知識を発見すること」と定義する．知識創発のための解釈を積み重ねるためには，テキストデータの分析時に試行錯誤を繰り返す必要があり，この試行錯誤のバリエーションを用意することで，得られる結果の幅が広がると想定される．

12 名の大学生，大学院生に対して，26 段落 (3574 字) からなる作家とアナウンサーの対談テキストを用いて，テキストの分析を行ってもらう実験を行った．被験者にはまず，文章を読んでもらった後，文章全体についてのデータ（グラフ）をもとに，その文章全体について言える事柄を挙げてもらった．その後，指定した個々の話題（単語）についてのデータ，また 2 つの話題の比較データをもとに，話題間の特徴を比較してもらい，再び文章全体について言える事柄を挙げてもらった．

*6 参加者は全員未読の作品を選択した．

それぞれの回答は上限を 5 個としており，前者で合計 44，後者で合計 48 の事柄が挙げられた．前者の中には，部分的な話題に言及した事柄は 4 つ (2 名) しかなかったのに対して，後者の中には，22(10 名) が部分的な話題に言及するとともに，そのほとんどが複数の話題を比較した内容を挙げていた．このことから，実装した内容が，複数の話題に関する新たな事実の列挙，ならびにそれらの解釈を支援する効果が確認された．

4. 関連研究

TETDM で想定している利用者はソフトウェアの使用に熟練した人だけでなく，初心者も含まれる．初心者でも簡単に利用でき，情報抽出，理解，および新しい知識の発見が可能な統合環境の構築を目標としている．

TETDM では利用者に対して TETDM の使用の理解を促すチュートリアルを用意している．加えて，チュートリアルの効果を確認できる事前テストと事後テストを用意している．このテストを受けることによって，チュートリアルで学習した効果を，利用者自身が確信できる．チュートリアルと事前，事後テストは，ソフトウェアの学習において初心者の好奇心や理解する喜びを定着させる効果がある [大槻 88]．

TETDM では 1 つのマイニング結果を 1 つの可視化だけで眺めるのではなく，複数の可視化で眺めたり，反対に複数のマイニング結果を 1 つの可視化で眺めることもできる．都市のアクティビティを表す指標を可視化し，都市の分析を支援するシステム [Brodbeck 03] や，1 つのデータに対して複数の可視化結果を表示し，1 つの可視化結果の中でデータを修正するともう一方の出力が修正したデータを反映して再出力されるシステム [Siirtola 03] では連動の機能が実装されている．連動が実現されているとデータを少し修正して結果を見てという行為を繰り返し行えるので，分析が容易になる．

インタラクティブに操作ができることと，出力結果が連動して変化することを組み合わせたことで，データの特徴にユーザが気づきやすくなる効果が生まれる．既存システムの中にも対話的にクラスタリングを進めて行けるシステムが存在する [井上 09] が，連動と組み合わせたものは少ない．

TETDM では過去に利用した人のログを残すような機能は現時点では存在しないが，将来的には機能を設けて利用者のモジュール利用に役立てて行きたい．使用すべきツールを思い出すことに時間がかかっていては分析がスムーズに進まない．過去の利用経験を自分が，また他人が追体験できるようにしておくと，過去に利用した人がその中で得た知識や経験を共有することができる [Nilsson 04]．動画の閲覧履歴を記録しておき，他の人の動画の見方で動画を見ることができるシステムなども提案されている [高嶋 08]．過去の操作ログを動画として残すだけでも価値があるが，利用者が見ていた視線の情報なども併せて記録することによって，TETDM の利用スキルを向上させていくことが可能となる．

5. 結論

本稿では，TETDM の利用者支援に関する実装と実践について述べた．TETDM の初心者の利用に際して，そのスムーズな利用を促す指針を示すとともに，これまでに実装したチュートリアルとその効果を確認した．また，テキストマイニングの実践例を示し，TETDM が提供する環境が幅広い目的に利用可能となることを示した．最後に，TETDM を用いた知識創発に向けた枠組みを示し，現在までに実装した内容とその評価結果について述べた．

今後は，示されたデザイン指針，ならびに知識創発に向けた枠組みをもとに，利用者支援に関わる実装を続けるとともに，モジュールを作成する開発者支援についても，具体的に実装と評価を行っていきたいと考えている．

参考文献

[Brodbeck 03] D. Brodbeck and L. Girardin, Design study: Using multiple coordinated views to analyze geo-referenced high-dimensional datasets, Proceedings of the Conference on Coordinated and Multiple Views In Exploratory Visualization, pp.104 – 111 (2003).

[井上 09] 井上悦子, 吉廣卓哉, 中川優：大規模クラスタリング結果のグラフによるインタラクティブな可視化手法, 電子情報通信学会論文誌, Vol.J92-D, No.3, pp.351 – 360 (2009).

[西原 09] 西原陽子, 佐藤圭太, 砂山渡：光と影を用いたテキストのテーマ関連度の可視化, 人工知能学会論文誌, Vol.24, No.6, pp.480 – 488 (2009).

[Nilsson 04] Nilsson, M., Drugge, M., and Parnes, P.: Sharing experience and knowledge with wearable computers, Proceedings of the Workshop on Memory and Sharing of Experiences (2004).

[大槻 88] 大槻説乎, 山本米雄：知的 CAI のパラダイムと実現環境, 情報処理, Vol.29, No.11, pp.1255 – 1265 (1988).

[Siirtola 03] H. Siirtola, Combining parallel coordinates with the reorderable matrix, Proceedings of the International Conference on Coordinated & Multiple Views in Exploratory Visualization, pp.63 – 74 (2003).

[砂山 14a] 砂山渡, 高間康史, 西原陽子, 梶並知記, 串間宗夫, 徳永秀和：統合環境 TETDM を用いたマイニングツールの開発と利用の実践, 人工知能学会論文誌, Vol.29, No.1, pp.100–112 (2014).

[高嶋 08] 高嶋章雄, 田中譲：習慣的な動画閲覧行動の再利用による動画閲覧経験の拡張, 情報処理学会論文誌, Vol.49, No.7, pp.2589—2597 (2008).

[井須 14] 井須弘恵，大塚直也，松下光範：探索的情報アクセスの支援に向けた TETDM インターフェースの改良に関する基礎検討，第 6 回人工知能学会インタラクティブ情報アクセスと可視化マイニング研究会 (2014).

[中垣内 14] 中垣内李菜, 川本佳代, 砂山渡:統合環境 TETDM を用いたテキストマイニング初心者のスキル獲得支援，第 28 回人工知能学会全国大会, 1H5-NFC-01b-3 (2014).

[北村 14] 北村侑也，高間康史：TETDM を用いたインタラクティブクラスタリングシステムの構築，第 28 回人工知能学会全国大会, 1H5-NFC-01b-1 (2014).

[徳永 14] 徳永秀和:TETDM による Exploratory Search の評価実験，第 28 回人工知能学会全国大会, 1H5-NFC-01b-5 (2014).

[砂山 14b] 砂山渡：統合環境 TETDM を用いた知識創発支援の枠組み，第 6 回人工知能学会インタラクティブ情報アクセスと可視化マイニング研究会 (2014).